Classification of Phonocardiogram Recordings Using Vision Transformer Architecture

Joonyeob Kim[1], Gibeom Park[1], Bongwon Suh[1,2]

[1]Department of Intelligence and Information, Seoul National University, Seoul, South Korea
[2]Research Institute for Convergence Science, Seoul, South Korea

Abstract

We participated in the George B. Moody Challenge 2022 to make a model which detects the presence or absence of murmurs from multiple heart sound recordings from multiple auscultation locations, as well as detecting the clinical outcomes from phonocardiogram (PCG) well. Our team, HCCL, developed a model with a visual approach for deriving a high-performance model. The model converts heart sound signals into spectrograms without requiring resampling or signal filtering. The result shows a weighted accuracy score of 0.69 (ranked 21th out of 40 teams) for the murmur detection classification on the hidden test data. For the clinical outcome identification task on the hidden test data, it shows a Challenge cost score of 11943 (ranked 6th out of 39 teams)

1. Introduction

Heart conditions and heart disease usually can be diagnosed through echocardiography or electrocardiogram. However, in underdeveloped countries, cardiologists and equipment are scarce, so they can only use PCG (phonocardiograms) which can be easily measured with a stethoscope. Nevertheless, interpreting sounds requires experts to interpret the results, so developing automated approaches for detecting abnormal heart function from multi-location PCG recordings of heart sounds at The George B. Moody Challenge 2022 will help diagnose and treat heart conditions [1–3].

In recent deep learning research, applying Transformer architecture [4] has shown excellent results for natural language processing tasks such as BERT [5] as well as for computer vision tasks. In the field of computer vision, image classification models based on Transformer have achieved better performances than traditional CNN-mixed architectures. In addition, the self-attention technique, which is an application of the Transformer model, could generate visualizations that could help understand the results. ViT (Vision Transformer) [6] is one of image classification models based on Transformer. The model splits an image into multiple patches and uses the embedding of individual patches and exhibited excellent performances.

Although deep learning models have been used for medical data analysis in many studies, there has been no case of modeling PCG recordings of heart sounds with a visual transformer so far. Inspired by this idea, we aim to explore the application of ViT architecture to classify PCG recordings with heart murmur patterns. Our final method is composed of time masking, signal segmentation without overlap, and converting a PCG into multiple spectrogram patches. In addition, the self-attention part of the model generates the attention map which could help understand the model's results. This approach allows us to distinguish heart murmurs recorded in PCG with higher accuracy as well as visualize the attention of the model on a spectrogram.

2. Methods

Our main research goal is to derive a high performance classifier based on the ViT model, which is no need for complex pre-processing of electrocardiogram [7] signals as well as finding visual features. Our method starts by converting heart sound to spectrogram which visualize the representation of the spectrum of frequencies of a signal varies with time (Fig 1.(a)). Using the spectrogram, we perform additional training using the pre-trained model which was trained on ImageNet-21K datasets. It allows the model to learn various features quickly and even with a small amount of data [6]. In addition, to address the challenge of small-sized training dataset, we used data augmentation method suitable for the model to avoid overfitting. Furthermore, we included demographic information such as gender and age into the classifier.

2.1. Dataset

The dataset of the George B. Moody Challenge 2022 contains one or more heart sound recordings for 1568 patients and routine demographic information about the pa-

ISSN: 2325-887X DOI: 10.22489/CinC.2022.084

tients [2, 3]. We do not use any extra data other than the training dataset provided by the George B. Moody Challenge 2022.

2.2. Pre-processing

The goal of preprocessing is to 1) turn the PCG recordings into spectrograms and 2) divide them into patches, so that they can be provided into our ViT module. We perform the following three pre-processing. During the process, we did not exclude or relabel the training data because our team has no experts on the given data. Also, no signal filtering or resampling was performed.

Since the pre-trained ViT model accepts 224×224, we create spectrogram into 224×224 pixels to avoid the loss of information from resizing spectrograms. For the same reason, the same patch size as the pre-trained ViT, 16×16, is used.

Signal segmentation without overlap

We processed the PCG recordings and created spectrograms of identical size. According to the paper of Raza, A. [8], in the case of PCG, the best performance was achieved when 12.5s was used as the input of the deep learning model. Based on the observation, we also split the original sound signal into 12.5s segments. When the length of a PCG recording is shorter than 12.5s, the image is padded by black pixels, giving the signal the same effect as zero padding. If the PCG is longer than 12.5 seconds, it crops out the beginning and end, then creates as many 12.5s segments as it can make. For example, if the PCG is 30 second long, the first and the last 2.5s are cropped and two 12.5s segments were created. When converting each segment to spectrogram, the following conversion parameters were used - window size of 0.11s and the overlap ratio of 0.5. The process produces spectrograms of a 224×224 matrix. No resampling or filtering was performed to minimize the loss of original data. The model performed better with the segmentation strategy. It might be due to the data augmentation effect.

Data Augmentation

Since the proposed model utilized only data provided in the competition, the size of data was insufficient and overfitting occurred. Thus, we tried to resolve the issue with data augmentation. Commonly used image augmentation methods are scaling, cropping, flipping, rotation, contrast, and saturation. However, the augmentation methods could compromise the important information in the spectrogram. Therefore, we applied the SpecAugment which is an augmentation method that works on the spectrogram of input signal [9]. SpecAugment includes various augmentation methods such as frequency masking, time masking, fade in and fade out. In the hidden validation data, time masking showed high performance than any other augmentation.

Demographic information

Our team used age, sex, height, weights, pregnancy information of the Challenge dataset. In the case of categorical data, it was converted using a label encoder, and each encoder was saved and used for the conversion of test data. For missing values, they were classified and converted into a new class. In the case of numeric data, it was transformed using the MinMax scaler, and each scaler was saved and used for the transformation of the test data. For a missing value, the mode was used for filling up the missing value. These values were also used for the test dataset.

2.3. Model

We used the ViT model of which patch size is 16×16 pixel and image size is 224×224 pixel. The model uses the pre-trained model of ImageNet-21K (14 million images, 21843 classes) [6]. The classifier part was modified to add demographic information to the model. We the following hyperparameters for the murmur task and clinical outcome task, an initial learning rate of 0.0001 with AdamW and LambdaLR scheduler. In the case of murmur classification, the batch size was 64, and saving steps and evaluation steps were 100. However, we used the batch size of 32, and saving steps and evaluation steps were 50 for outcome classification. The larger the batch size of ViT, the better the performance, so the maximum possible value in the test environment was used, and in the case of step, the optimal value was found experimentally.

Vision Transformer

ViT splits the image into small patches and performs image classification. The ViT model shows higher performance with less data during fine-tuning than CNN-based models. Furthermore, it has the advantage of self-attention that can generate an attention map that could help understand the model's results. In ViT, attention map and mask are calculated by Attention Rollout [10]. To compute the attention mask and map, the attention from the encoder is averaged over the head, and recursively multiply the weights of all layers. Multiply the attention mask as if applying a filter to the spectrogram image to get an attention map. In the attention mask (Fig 1.(b)), the yellow parts which are bright than the other parts are the attention part. It mainly paid attention to low frequencies, but occasionally attended to high frequencies, and this information might helpful for classification.

We used the pre-trained model of ImageNet-21K because the model is hard to perform well without pretraining since a large amount of data is often required for Transformer.

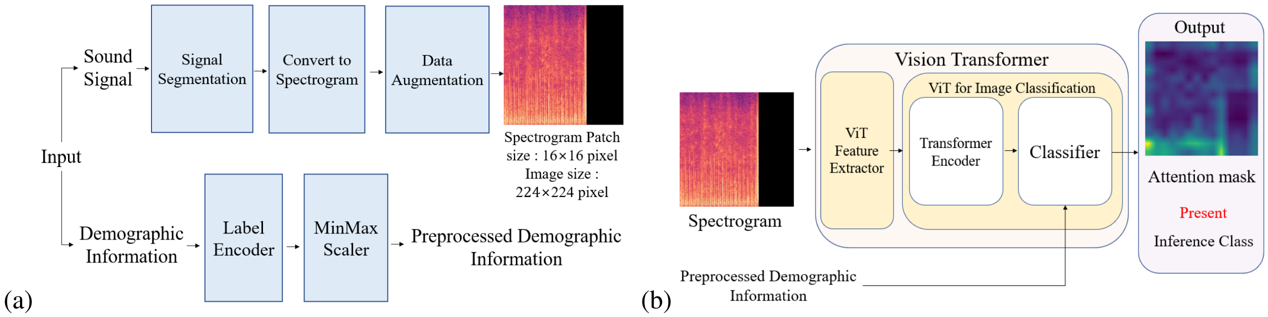

Figure 1. (a) Pre-processing overview. (b) Model overview.

Model architecture

Our model is as shown in Fig 1.(b). Pre-processed heart sound signals are converted into images, and it is input to ViT Feature extractor. As the output of ViT feature extractor, the encoded feature is output in batch units, and it is input to the Transformer encoder of ViT for Image Classification. The input features go through a total of 12 ViT layers, and the pre-processed demographic information features are concatenated into the output. The concatenated feature is input to the last layer, the classifier. Finally, the class is learned and inference is performed.

Weighted categorical cross-entropy loss

Cross-entropy is used to measure the difference between different probability distributions and is used as a loss function for classification. Since the class of the the Challenge dataset is imbalanced and each class has a different degree of significance, a weighted categorical cross-entropy loss is applied. The weighted categorical cross-entropy loss is shown as (1):

$$\text{Ł} = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} y_m^k \times w_k \times \log\left(h_\theta(x_m, k)\right) \qquad (1)$$

where M is number of training examples, K is number of classes (3 at murmur, 2 at outcome), y_m^k is target label for class k at each training example m, w_k is weight (1, 5, 3 for Absent, Present, Unknown class at murmur, 1.2, 1 for Abnormal, Normal at outcome), h_θ is model with weights θ, x_m is input for training example m [11]. In general, the methods for determining weights: using all the same weights, using a reciprocal of the class ratio, or using a ratio of cost function weights. In the case of the murmur task, the weight ratio of 1:5:3 which is the same as the ratio of the challenge murmur cost function was the best performance. In the clinical outcome task, a weight of 1.2:1 was selected by fine adjustment, which is the best performance hyperparameter empirically in the hidden validation set.

Training	Validation	Test	Ranking
0.936	0.671	0.69	21/40

Table 1. Weighted accuracy metric scores (official Challenge score) for our final selected entry (team HCCL) for the murmur detection task, including the ranking of our team on the hidden test set. We used 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

Training	Validation	Test	Ranking
10615	9903	11943	6/39

Table 2. Cost metric scores (official Challenge score) for our final selected entry (team HCCL) for the clinical outcome identification task, including the ranking of our team on the hidden test set. We used 5-fold cross validation on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

3. Results

The results of our model on the murmur detection task for the challenge data set are shown in Table 1, and the results of the clinical outcome identification task are shown in Table 2. Our model placed 21st out of 40 teams and 6th out of 39 teams respectively.

The results of an ablation study for the murmur detection and the clinical outcome identification are reported in Table 3 and Table 4. Our final model contains signal segmentation, data augmentation, and demographic data as well as pre-trained weights. The majority of preprocessing techniques enhanced the performance. In the both tasks, performance was most influenced by using the pre-trained model on the training dataset. Next, augmentation contributed to improving performance. On the other hand, signal segmentation and demographic data had negligible effects on the training dataset.

Methods	Weighted accuracy
Final model without Segmentation	0.751 ± 0.025
Final model without Pretrained	0.661 ± 0.016
Final model without Augmentation	0.704 ± 0.042
Final model without Demographic	0.744 ± 0.025
Final model	0.759 ± 0.030

Table 3. Results of our final model's ablation study of the murmur detection task using 5-fold cross validation.

Methods	Challenge cost
Final model without Segmentation	14800 ± 4380
Final model without Pretrained	23514 ± 4976
Final model without Augmentation	13726 ± 912
Final model without Demographic	14283 ± 1821
Final model	13091 ± 1183

Table 4. Results of our final model's ablation study of the clinical outcome identification task using 5-fold cross validation.

4. Discussions

We studied a model that can detect murmur and pathological outcome by applying the visual approach with the ViT model to PCG modeling. Contrary to expectations, in the murmur detection task, the score of the hidden validation and test data was lower than the training data. In the outcome identification task, the score of the training data was higher than the hidden test data, but lower than the hidden validation data. It seems that the model is overfitted to the training dataset, and we expect to get better results by applying the regularization skills. In the case of signal segmentation, the training data showed a score improvement of 3.3%, but the hidden validation data showed a score improvement of 13.3%, which showed that signal segmentation was effective in the PCG murmur detection for the ViT model.

Due to our team's limited expertise on the subject, there might be a better optimized pre-processing technique. Although we heard PCG, detailed information could not be obtained due to a lack of pathological understanding of the data. Potential anomalies could have not been detected in the demographic information. According to Table 3, it is confirmed that the model is overfitted to the training dataset, and if more regularization techniques are applied to the our murmur task model, a better performing model could have been generated. If exclusion criteria or relabeling is performed through an expert on data, or additional data is utilized, the model could have performed better.

One potential benefit of the proposed method is that it allows us to examine an attention mask giving opportunities what part of PCG are informative for each pathological condition. We believe that this could open up opportunities for experts to understand, interpret, and improve the model's findings.

Acknowledgments

This work was in part supported by the Research Institute for Convergence Science.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, Physiotoolkit, and Physionet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220.

[2] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[3] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[4] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems 2017;30.

[5] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv181004805 2018;.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv201011929 2020;.

[7] Naz M, Shah JH, Khan MA, Sharif M, Raza M, Damaševičius R. From ecg signals to images: a transformation based approach for deep learning. PeerJ Computer Science 2021;7:e386.

[8] Raza A, Mehmood A, Ullah S, Ahmad M, Choi GS, On BW. Heartbeat sound signal classification using deep learning. Sensors 2019;19(21):4819.

[9] Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, et al. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv190408779 2019;.

[10] Abnar S, Zuidema W. Quantifying attention flow in transformers. arXiv preprint arXiv200500928 2020;.

[11] Ho Y, Wookey S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access 2019;8:4806–4813.

Address for correspondence:

Bongwon Suh
Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea
bongwon@snu.ac.kr